Time Warping of Audio Signals

Siome Goldenstein
VAST Lab. – University of Pennsylvania
200 South 33rd Street,
Philadelphia, PA, USA 19104
siome@graphics.cis.upenn.edu

Jonas Gomes
IMPA–Instituto de Matemática Pura e Aplicada
Estrada Dona Castorina 110,
Rio de Janeiro, RJ, Brasil 22460
jonas@visgraf.impa.br

June 18, 1999

Abstract

This paper describes a technique to obtain a time dilation or contraction of an audio signal. Different Computer Graphics applications can take advantage of this technique. In real-time networked VR applications, such as teleconference or games, audio might be transmited independently from the rest of the data, These different signals arrive asynchronously and need to be somehow resynchronized on the fly. In animation, it can help to automatically fit and merge pre-recorded sound samples to special timed events. It also makes it easier to accomplish special effects like lip-sync for dubbing or changing the voice of an animated character. Our technique tries to eliminate distortions by the replication of the original signal frequencies. Malvar wavelets are used to avoid clicking between segment transitions.

Keywords: Audio Transformation, Wavelets, Local-Cosine Transform, Dynamical Time Warping, Time-Frequency Manipulation, Audio and Video Synchronization, Speech Recognition.

1 Introduction

Audio information is becoming an important tool for Computer Graphics. Many applications either use audio to augment their effects or interaction or are intrinsically coupled with some audio information. In Virtual Reality applications audio effects synchronized with particular events can convey much more information. For example, the sound of a ball hitting a surface can give the user some hints about these objects' properties.

Since audio has different properties and characteristics from other standard computer graphics objects, real-time networked applications (like VR, interactive games or teleconferencing), can choose to transmit and process it separately from the rest of the data. These different real-time data streams will arrive possibly out of synchronization, and somehow must be correctly realigned. One approach to solve this is to constantly shrink the audio stream when it is late, and expand it when it is ahead. In some lip-sync techniques video is generated or altered in order to match audio [4, 16]. In the particular case where the dubbing process just changes the speakers voice and not the contents, the right facial movement and expressions are already there, just in a different timing. In this situation it might be easier just to alter the lengths of segments of audio instead of manipulating the video. Video editing and post-processing can also take advantage of shrinking and stretching voice and audio signals to match specific video cuts.

In this paper we study the special case of *time warping*, where the object is a digital audio signal and the deforming operation will be applied over its geometric support: an interval of time.

The ultimate goal of this transformation is to change the duration of the digital audio signal without perceptual information change. This means that if the operation is applied to a human voice recording, the result would be recognizable as the same speaker pronouncing exactly the

same content, but at a different "speed" (slower or faster).

Generic object warping operations have been used in many applications related to computer graphics (animation, multimedia, etc.) The ability to perform arbitrary changes and deformations of graphical objects allows many interesting and useful applications. The operation of warping and morphing has been studied in the overall setup of graphical objects in [9, 8]. This operation acts on both the shape and attributes of the graphical objects.

The organization of the paper is as follows: Section 2 briefly reviews previous works in this area; Section 3 discusses time warping in the time domain; Section 4 studies time-frequency representation of sound signals; Section 5 describes our method of applying a time-warping transformation in the frequency domain; Section 6 gives an example and discusses some applications.

2 Previous Work

The first attempt to develop a time warping transformation is the technique of Dynamic Time Warping (DTW) described in [17] and [7]. This method "extends" the signal length by replication or/and elimination of samples. This approach has been in use for size matching reference and input frames for speech recognition.

In computational music, there is a long list of works dealing with transformations; some of them take the same direction we will describe below. In [3] and [2], the authors use a Gabor time $\times$ frequency representation and take great care about the "legality" of its transformation results. Special attention is taken to the phase component of the image. In [13] and [14] wavelet transforms are the tool for the transformations. In these cases, the working tools are continuous-time transforms, implying in harder, and sometimes not so efficient implementations.

There is some interesting work that is being developed in sound processing and manipulation, [1].

Our approach is novel in two ways. First we use a discrete transform, which allows fast and efficient implementation. Second, by the choice of this transform: local cosine basis, which is a real orthonormal transformation (perfect reconstruction) that achieves window overlapping and dismisses the care in the "phase" component.

3 Time Warping in Time Domain

The simplest way to describe an audio signal is to use a temporal representation. Mathematically, an audio signal is a function that measures the air pressure variation over time. This can be mapped in a straightforwardly way to the graphical object notation described in [9]. The time interval where the sound occurs is the geometric support, and the air pressure will be its attribute function.

Given an audio signal $f\colon [a,b] \to \mathbb{R}$, where $[a,b]$ is the time interval, and a function $T\colon [a,b] \to [c,d]$, we can obtain a new audio signal $g\colon [c,d] \to \mathbb{R}$, where $c = T(a)$, $d = T(b)$ and $g(t) = f(T(t))$. Mathematically this is just a reparameterization of the time interval. When T is an affine transformation, we have

$$T(t) = \frac{t-a}{b-a}(d-c) + c. \tag{1}$$

In this way a time dilation or a time contraction are described. From now on, we shall suppose that time dilation or contraction will be done using affine mappings as in (1). A simple example of time dilation is illustrated in Figure 1 for a sinusoidal signal.

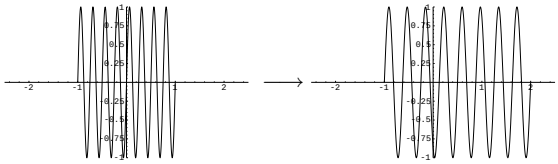

Figure 1: Expansion in time domain.

The Dynamic Time Warping method [17, 7] implements a transformation similar to this, working with the discrete representation of the audio.

It is well known that a more robust approach is to reconstruct the audio signal, apply the time warping, and resample it again. In fact, since the time transformation might introduce high frequencies, it is advisable to filter some high frequencies before resampling. This is a well-known operation called *sampling rate conversion* in the area of signal processing, and *resampling* in the context of warping and morphing [10].

Since the human auditory perception is related to the oscillations of the air pressure, it is essential to the analysis

of the spectral effects of the transformation. From Fourier Analysis we have:

$$f(at) \leftrightarrow \frac{1}{|a|}\widehat{f}(\frac{\omega}{a}),$$

where $\widehat{f}(\omega) = \mathcal{F}\{f(t)\}(\omega)$.

Therefore, it becomes clear that a simple time compression/dilation will also change the frequencies components of the original signal, creating undesirable perceptual distortions. In fact, a naive time dilation or compression, perceptually produces the effect of playing an old vinyl long-play played at a wrong rotation.

To accomplish time-warping operations correctly, we must take into account the close relationship between time and frequency in an audio signal.

4 Time $\times$ Frequency Representation

A logical improvement is to use a better representation of the audio signal, describing both time and spectral localization characteristics. There is vast signal processing literature concerning this theory, such as [12, 5, 6].

The two most common linear time $\times$ frequency representation are the Window Fourier Transform (WFT), also called Short Time Fourier Transform, and the Wavelet Transform. Each has its pros and cons. We have chosen the *Local Cosine Transforms* [15]. This transform is also called *Malvar Wavelets*; for a detailed discussion, see [19].

The Malvar wavelets have two relevant advantages over the Window Fourier Transform:

1. It is a real transform based on the Discrete Cosine Transform-IV [18]. This avoids the need for special care of the phase component.

2. It is an orthogonal transform, although its windows (which do not need to have the same size) overlap.

The overlapping is responsible for the elimination (or considerable reduction) of the undesirable clicking, that usually appears after synthesis and manipulation using the WFT representations (due to discontinuities in the boundaries of adjoining windows).

The forward transform is accomplished in two steps: first a "folding" operation is done on each segment, which will in some sense add the neighbors' border information to each window; then, a normal DCT-IV is executed. This folding operation must be carefully projected such that an orthonormal transformation is achieved at the end. A whole family of this step is described in [19] and [11].

The inverse transform is also done in two steps: the normal DCT-IV (which is it own inverse), followed by the unfold operation.

Some of the different elements of this basis can be seen in Figure 2. Figure 3 illustrates the overlapping of different basis elements. They are carefully constructed such as to preserve orthonormality.

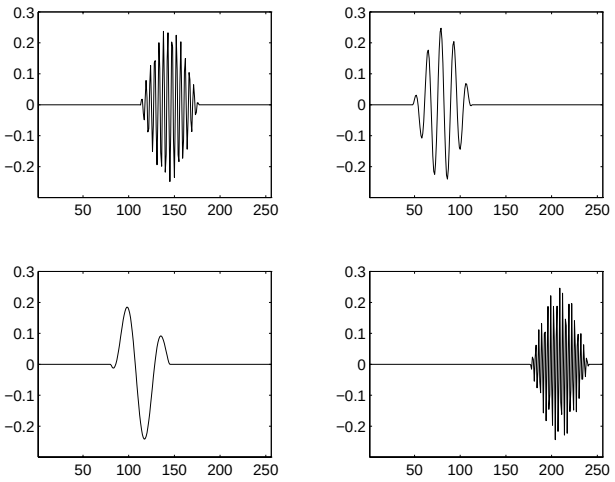

Figure 2: Four different elements of the basis.

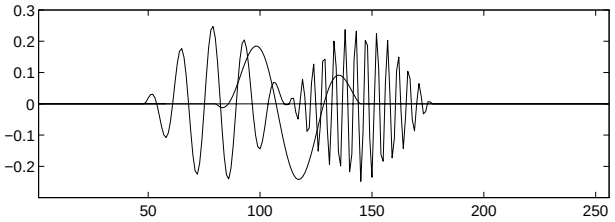

Figure 3: Three basis elements: Orthonormality with overlap.

5 Time Warping in Time $\times$ Frequency Domain

When an audio signal is represented in a time-frequency domain, it can be regarded as a continuous 2D image. The support is a rectangle whose horizontal axis represents the time support of the phenomenon, and whose vertical axis represents the frequency axis (from 0 to π). The bottom picture in Figure 4 shows the time $\times$ frequency representation of the audio signal shown on the top.

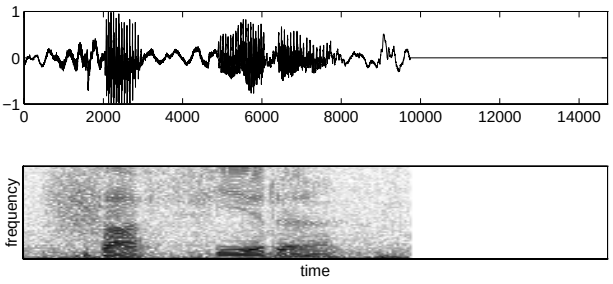

Figure 4: Time-frequency representation.

The time warping operation in the frequency domain uses an affine dilation on the time axis of the time-frequency representation (i.e., we scale the image on the time axis). We transform the audio signal, apply the time scaling (dilation or compression), and reconstruct the audio in the time domain using the inverse transform. This is illustrated in Figure 5: (a) shows the time-frequency representation; (b) shows the original audio signal in time domain; (c) shows the time dilation of the representation in (a), and (d) shows the reconstructed audio signal in time domain after the time warping operation of (c).

Since regions of the image represent the presence of certain frequency components in the time segment limited by its boundaries, its stretching is responsible for a replication of the oscillations (prolonging the phenomenon in the expansion case).

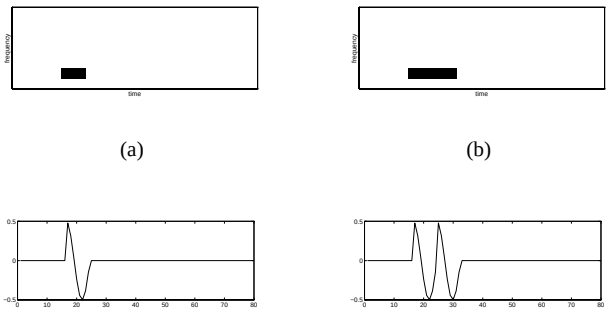

Figure 5: Time dilation in time-frequency domain.

6 Examples and Applications

Figure 6 shows a time dilation applied to a digital speech recording of the word “safira” (sapphire in Portuguese) at 8 kHz with 8 bits per sample with a expansion of 50% using a 512-sample window. In (a), we show the original audio in time domain; in (b) we show the time-frequency representation; in (c) we show the time dilation in time-frequency domain, and in (d) we show the reconstructed audio waveform in time domain.

Although at this scale, the time dilation only seems to have accomplished a simple time-stretching, this is not the truth. Figure 7 shows an amplified image of the time representation region between samples 2000 and 3000. Figure 7 (a) corresponds to the original sound, and Figure 7 (b) is the warped sound. Observe carefully that the number of “peaks” of the warped sound is twice that number in the original sound.

Note that some oscillations were not “stretched” (observe on Figure 6 the “slow” oscillations between samples 3000 and 4000). This undesirable result has to do with the windowing partitioning done underlying the whole process. It is impossible to stretch or compress oscillations that are not enclosed in one full window. Although one can always use a larger window, there is a certain price to be paid: according to the uncertainty principle, the larger the window, the more temporal localization is lost.

It was observed that for large expansions (around 100%)

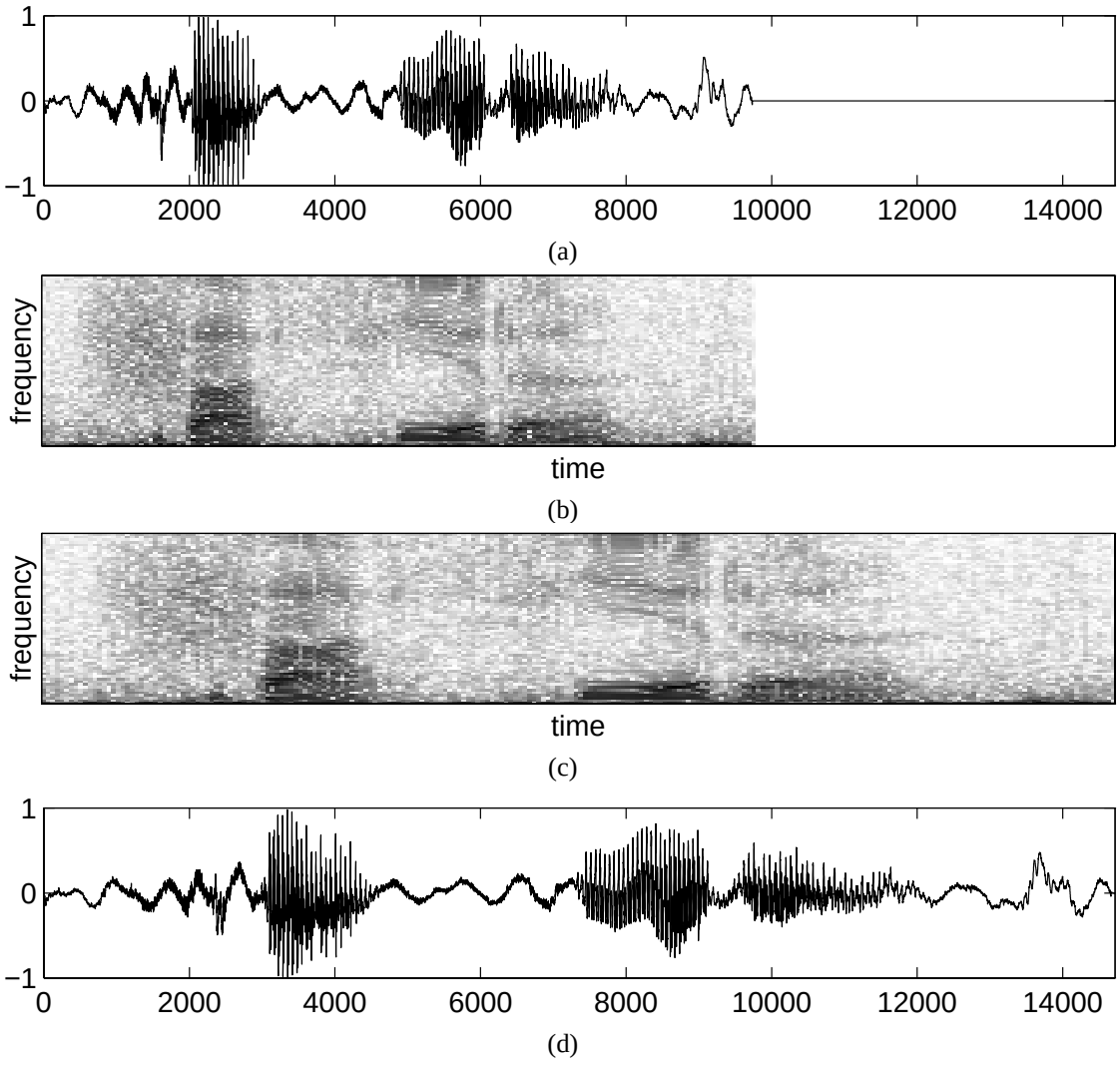

Figure 6: Time dilation of the word “safira”.

of speech signals the results start to present some severe perceptual problems. This is because the warping process was being applied uniformly, and in real speech some transient portions have always the same temporal length, no matter how slowly the speaker is talking.

For a correct result, it would be necessary to accomplish a temporal segmentation of the speech signal, identify which segments are “allowed” to be warped and how much, and after that glue everything together again. Unfortunately, neither of these tasks is easy.

In virtual reality, interactive games and real-time teleconference applications audio is transmitted apart from the rest of data, and sometimes they lose synchronism. One possible way to achieve the correct correspondence again is to modify the sound track slightly, inside a much more complex control system. This can be easily done using

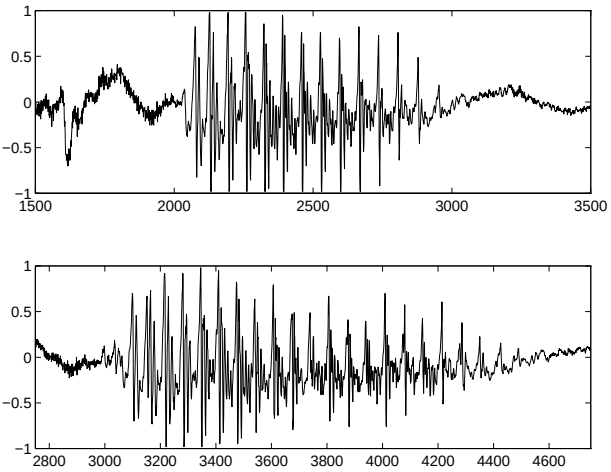

Figure 7: Waveform detail of time dilation.

our method described here.

Other typical applications of this transformation are non-linear editing of audio and video sequences and film dubbing (lip-sync). In these applications sometimes the speech part of the audio is recorded completely independently from the video. This makes a posteriori synchronization of audio and video tracks a really difficult task. By slight adjustments of the length of different segments of the audio track, it is then possible to match image with sound information, without significant distortion.

Another application is in speech recognition and speaker identification systems, where it is important to match "segments" of the signal with some reference frames. Unfortunately the real segment will never have the same "length" as the reference, so some kind of time warping has to be done. This is mostly done with DTW. Eventually, the distortion caused by the warping process can compromise the overall performance of the system.

7 Conclusion and Future Work

We have presented a technique to compute time warping of audio signals. The technique works on a time-frequency representation of the audio signal, using Malvar wavelets.

Our method does not rely on any audio or speech specific attribute (like formants or specific sampling rate), it uses and replicates periodic information. We are currently studying how to apply this technique for animation and motion capture signals that have these periodic characteristics.

8 Acknowledgments

This research has been developed in the laboratory of VISGRAF project at IMPA. This project is sponsored by CNPq, CAPES, FAPERJ, FINEP, and IBM Brasil.

References

[1] http://www.iua.upf.es/eng/recerca/mit/sms/docs/software.html.

[2] D. Arfib. Analysis, transformation, and resynthesis of musical sounds with the help of time-frequency representation. In G. D. Poli, A. Piccialli, and C. Roads, editors, *Representations of Musical Signals*, pages 87–118. MIT Press, 1991.

[3] D. Arfib and N. Delprat. Musical transformations using the modification of time-frequency images. *Computer Music Journal*, 17(2):66–72, 1993. (Sound examples on soundsheet with 13(1) 1989).

[4] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *SIGGRAPH 97 Conference Proceedings*, pages 353–360, 1997.

[5] L. Cohen. Time-Frequency Distributions - A Review. *Proceedings of the IEEE*, 77(7):941–981, 1989.

[6] I. Daubechies. The Wavelet Transform, Time-Frequency Localization and Signal Analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, September 1990.

[7] J. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete Time Processing of Speech Signals*. Prentice Hall, 1995.

[8] J. Gomes, B. Costa, L. Darsa, and L. Velho. Graphical objects. *The Visual Computer*, 12(6):269–282, 1996.

[9] J. Gomes, B. Costa, L. Darsa, and L. Velho. *Warping and Morphing of Graphical Objects*. SIGGRAPH '97 Course Notes. Los Angeles., 1997.

[10] P. Heckbert. Fundamentals of texture mapping and image warping. Master thesis (technical report no. ucb/csd 89/516), University of California, Berkeley, 1989.

[11] E. Hernández and G. Weiss. *A First Course on Wavelets*. CRC Press, 1996.

[12] F. Hlawatsch and G. G. Boudreaux-Bartels. Linear and Quadratic Time-Frequency Signal Representation. *IEEE SP Magazine*, pages 21–67, April 1992.

[13] R. Kronland-Martinet. The use of the wavelet transform for the analysis, synthesis and processing of speech and music sounds. *Computer Music Journal*, 12(4):11–20, 1988. (Sound examples on soundsheet with 13(1) 1989).

[14] R. Kronland-Martinet and A. Grossmann. Application of time-frequency and time-scale methods (wavelet transforms) to the analysis, synthesis, and transformation of natural sounds. In G. D. Poli, A. Piccialli, and C. Roads, editors, *Representations of Musical Signals*, pages 45–85. MIT Press, 1991.

[15] H. S. Malvar. *Signal Processing with Lapped Transforms*. Artech House, 1992.

[16] F. Parke and K. Waters. *Computer Facial Animation*. AK Peters, 1994.

[17] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall Inc, 1978.

[18] K. R. Rao and P. Yip. *Discrete Cosine Transform : Algorithms, Advantages, and Applications*. Academic Press, 1990.

[19] M. V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. A. K. Peters, Wellesley, MA, 1994.